Video Person Re-Identification for Wide Area Tracking based on Recurrent Neural Networks

Video Person Re-Identification for Wide Area Tracking based on Recurrent Neural Networks

Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller,

***Abstract*—In this paper we propose a video-based person re-identification system for wide area tracking based on a recurrent neural network architecture. Given short video sequences of a person, generated by a tracking algorithm, our video re-identification algorithm links these tracklets in full trajectories across a network of non-overlapping cameras in an open-world scenario. In our system, features are first extracted from each frame using a convolutional neural network. Then, a recurrent layer combines information across time-steps. The features from all time-steps are finally combined using temporal pooling to give an overall appearance feature for the complete sequence. Our system is trained to perform re-identification using a Siamese network architecture. Experiments are conducted on the iLIDS-VID and PRID-2011 video re-identification datasets as well as in the DukeMTMC multi-camera tracking dataset.**

***Index Terms*—Deep Learning, Re-identification, Recurrent Neural Networks, Wide Area Tracking.**

I. Introduction

In recent decades, more and more cameras have been installed in public and private spaces for monitoring activities and behaviours. While some computer vision systems have been deployed to perform automated analysis and video surveillance [49], [15], such deployment has been scarce, as few systems allow for automatic large scale wide-area monitoring [44]. This is due to a number of difficult problems that must be overcome in order to extend video surveillance from a single camera to a large network of cameras, including: different camera configuration, different appearance of targets between cameras, unknown camera layout and unknown building topology, to name but a few [7], [58].

Given the above context, this paper addresses the problem of wide area tracking with the aim of tackling the aforementioned challenges. We define wide-area tracking as the capability to track every subject of interest through a camera network, where the cameras have non-overlapping fields of view and are distributed over an unknown and arbitrary layout. By reformulating the wide area tracking problem as a re-identification problem, association between the tracks of people moving between cameras can be established regardless of changes of in the person's appearance caused by viewpoint and camera configurations. Our proposed re-identification system can be used in conjunction with a multi-target tracking framework to associate tracklets between non-overlapping cameras while incorporating spatial and temporal priors, should they be available, to mitigate the challenges of the unknown camera layout.

The video re-identification problem entails associating different tracks of a person as they move between non-overlapping cameras [11]. This occurs when the video of a person as seen in one camera must be matched against a gallery of videos captured by a different non-overlapping camera. The difficulties of person re-identification are due to large appearance changes caused by environmental and geometric variations as a person moves between cameras.

While re-identification has been extensively studied for single still images, the video-based re-identification problem has been much less studied despite its wide applicability in multi-person, multi-camera tracking. This could be due to several reasons including, the lack of large-scale video re-identification datasets available until recently [59], which has made it challenging to train effective video re-identification systems.

Using video data for re-identification has several important advantages over using single images. Firstly, the video setting is more natural since in realistic situations where person re-identification would be used, a person's image will normally captured by a video camera, which produces a sequence of images rather than a single frame. More importantly, video sequences contain person-specific temporal information related to motion and gait, which has been proved to be useful for differentiating between people or even being used as a soft biometric [45]. By recording and making use of this temporal information, the system may be able to disambiguate cases that would be difficult given only a single image. Finally, sequences of images contain a larger pool of samples for each target, where each sample may have slightly different poses, viewpoint, and background. This allows a better model of the person's appearance to be learned and makes training machine learning algorithms easier. This is specially relevant when using neural networks, which can be highly demanding in terms of the number of training samples required.

On the other hand, the use of video for re-identification also creates new challenges, such as comparing video sequences of arbitrary length and/or different frame-rates, the presence of unknown partial or full occlusions within the sequences, and the possibility of partial and cropped views rather than full-body images due to tracking inaccuracy during video sequence extraction.

The last problem can be mitigated by using an accurate multi-target tracker [30] to generate accurate tracklets. These tracklets can then be used by the video re-identification

Corresponding author: Niall McLaughlin. Centre for Secure Information Technologies (CSIT), Queen's University Belfast, UK, e-mail: n.mclaughlin@qub.ac.uk.

system, assuming that the multi-target tracker has discarded associations that are likely to be physically impossible.

The integration of re-identification and trackers also allows for the study of the open-world re-identification problem, where the probe subject may not necessarily be present in the gallery and vice versa. While dealing with this problem is crucial for real applications of wide-area tracking, it has been largely ignored in the re-identification literature.

II. Related work

While wide-area tracking and video re-identification have been scarcely studied, person re-identification for still images has been extensively investigated. Proposed methods for person re-identification broadly fall into two categories. The first of these aims to define, extract and use features that are both discriminative and invariant to environmental and view-point changes [29], [4], [8]. The second category employs supervised learning based methods that learn to automatically map the raw features into a new space with greater discriminative power [20], [18], [60]. Among the most promising and better performing techniques in this category, deep learning techniques [65], [6], [12] may be advantageous as they remove the need for hand-crafted features, and give state-of-art performance provided there is sufficient training data.

After features have been extracted, metric learning is widely used in person re-identification to learn a metric that emphasises inter-personal distance and de-emphasises intra-person distance. The learnt metric is used to make the final decision as to whether a person has been on the re-identified or not. Various methods have been proposed based on this idea such simple Mahalanobis distance, Relaxed Pairwise Learning (PRLM) [20], Large Margin Nearest-Neighbour (LMNN) [60], and Relevance Component Analysis (RCA) [2].

Only a few methods for video and/or multi-shot re-identification have been described in the literature. These include collecting interest-point descriptors over time [13], or training classifiers using features collected over multiple frames [43]. Furthermore, supervised learning based methods have also been used, such as learning a distance preserving low-dimensional manifold [5], or learning to map between the appearances in sequences by taking into account the differences between specific camera pairs [31]. Other approaches that explicitly model video include using a conditional random field (CRF) to ensure similar images in a video sequence receive similar labels [24], or extracting space-time features [27], [1] and then learning a ranking function that is robust to partially corrupted sequences [59].

Deep neural networks (DNN) have been applied in most areas of computer-vision [50], [28], [48], [55], and have largely replaced traditional computer vision pipelines based on hand-crafted features. Deep networks have also made an impact in image based person re-identification. Thus, DNNs have been used to learn ranking functions based on pairs [65], or triplets of images [6]. Specialised network architectures have been developed for directly comparing pairs of images. Network architectures such as the 'Siamese network' [12], are used to learn a direct mapping from the raw image pixels to a feature space where images from the same person are close, while images from different persons are as separated. The approach in [32] also allows comparing image pairs while taking into account deformation. Another DNN-based approach to re-identification uses auto-encoders to learn an invariant colour feature, whilst ignoring spatial features [57]. To address the large amount of training data inherent to DNNs, several approaches have been proposed for improving generalisation given limited training data [16], by using data augmentation [38], multi-task learning [52], [3] or unsupervised training [37], [9].

Within the video re-identification field a handful of recent architectures have been proposed to learn a feature representation for persons, based not only on spatial or appearance features, but also on some form of temporal information. Recurrent networks and temporal pooling [40], [62], [61], [64] have been shown to improve performance by modelling temporal information within a end-to-end trained DNN approach. A similar idea is presented in [34], where the spatial and temporal information is separated during learning using a double-stream recurrent network. However, all these approaches have been only tested in a closed-world scenario.

In this paper we propose a novel neural network architecture for wide area tracking and video based person re-identification. The network architecture uses appearance and motion information to extract a feature representation for each person that is invariant to illumination, angle of view, and pose. This means that given a short video clip of a person as seen in one camera, that person can be reacquired by a separate non-overlapping camera. The network receives as input a video of the colour and optical flow information from the cropped bounding box of a person, produced by an existing multi-target tracker in one camera. Temporal information is extracted using a recurrent layer and the information from the whole input video combined using temporal pooling to produce a fixed-size feature-representation for the whole input sequence. The network is trained using a contrastive loss function i.e., a Siamese architecture, to produce this invariant feature representation. Moreover, our network is tested in the context of open-world re-identification and integrated in a multi-target tracking framework to evaluate its potential for wide-area monitoring.

This paper builds on the work in [40] by investigating different recurrent network architectures such as standard RNN, Long Short Term Memory, Gated Recurrent Unit and configurations with and without residual connections, which may allow a better representation of the temporal information required for video re-identification. In addition, a more thorough evaluation, including results on the newly proposed MARS dataset [66], is performed, and a broader and more up-to-date comparison with the state of the art in video re-identification is provided. Furthermore, the video re-identification network is integrated within a single and multi-camera tracking framework to demonstrate its value for wide area tracking in realistic scenarios.

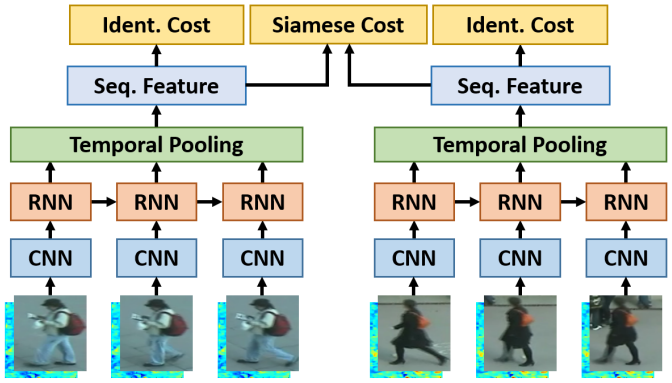

Fig. 1. Our proposed video-based re-identification system.

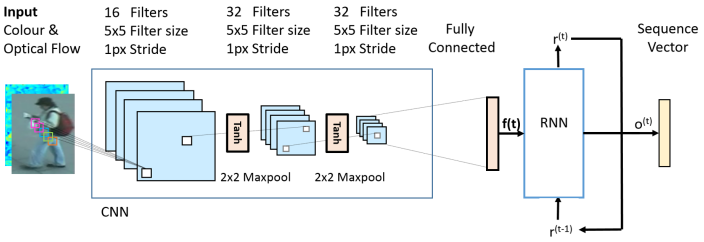

Fig. 2. The structure of our proposed CNN and recurrent layer, where $r^{(t)}$ is the RNN's state at time t and $o^{(t)}$ is the sequence vector output at time t. See Section III-B and Section III-C for details.

III. Video Re-identification Methodology

A diagram of our proposed feature extraction architecture is shown in Fig. 1. In our architecture each frame is first processed by a convolutional neural network to produce a feature vector representing the person's appearance at a particular instant in time. We then allow information to flow between time-steps by using a recurrent layer, before the outputs from all time-steps are combined using temporal pooling. Temporal pooling allows the network to summarise an arbitrarily long video sequence into a single feature vector, while the recurrent layer may allow the network to better exploit temporal information within the sequence, before the outputs from all time-steps are combined.

In order to train the feature extraction network to perform re-identification, we use a Siamese network architecture [12] as shown in Fig. 1. Given a pair of sequences from the same person, the Siamese architecture is trained to produce sequence feature vectors that are close in feature space, while given a pair of sequences from different persons, the network is trained to produce sequence feature vectors that are separated by a margin. This objective function mirrors the structure of the re-identification problem, where it must be decided whether two images depict the same person or not. In the following section we will explain each of the components of our proposed network in greater detail.

A. Input

The input to the convolutional network consists of both optical flow and colour channels. Thus, images are converted to the YUV colour space, before being passed to the network, and each colour channel is normalised to have zero mean and unit variance. Horizontal and vertical optical flow channels are calculated between each pair of frames using the Lucas-Kanade algorithm [36]. The optical flow channels are then normalised to fall within the range -1 to 1. As a consequence, the first layer of the neural network used five input channels, three for colour and two for optical flow.

While colour encodes details of a person's appearance and clothing, optical flow directly encodes short-term motion, which may include details of a person's gait as well as other motion cues. By using both colour and optical-flow together, the network should be better able to exploit short-term temporal information in order to improve re-identification accuracy compared with using colour alone.

B. Convolutional Network

As shown in Fig. 1, at each time-step the image is processed by a convolutional neural network (CNN). The CNN involves many individual processing steps, therefore for notational simplicity we refer to the complete CNN as a function, $f = C(x)$, that takes an image x as input and produces a vector f as output. In general, a CNN processes an image using a series of layers, where each individual layer is composed of convolution, pooling, and non-linear activation-function steps. In our case, we use max-pooling and the hyperbolic-tangent (Tanh) activation-function. Each layer of the convolutional network therefore performs the operation $C'(s^{(t)}) = Tanh(Maxpool(Conv(s^{(t)})))$, where in the first layer, the input, $s^{(t)}$, is the original image, and in deeper layers the input is the output feature maps from the previous layer of the CNN.

Let $\mathbf{s} = s^{(1)}...s^{(T)}$ be a video sequence, of length T, consisting of whole-body images of a person, where $s^{(t)}$ is the image at time t. Each image, $s^{(t)}$, is passed through the CNN to produce a vector, $f^{(t)} = C(s^{(t)})$, where $f^{(t)}$ is the vectorised representation of the CNN's final layer activation maps. The vector $f^{(t)}$ is then passed forward to the recurrent layer (see Section III-C), where it is projected into a low-dimensional feature-space and combined with information from previous time-steps. Note that the parameters of the CNN are shared across all time-steps meaning that each input frame is processed by the same feature-extraction network. Dropout [51] is used between the CNN and the recurrent layer in order to reduce over-fitting. Complete details of the CNN architecture are given Fig. 2.

C. Recurrent Network

Recurrent neural networks (RNN) address the problem of processing an arbitrarily long time-series using a neural network, which can be problematic for standard architectures with a fixed number of input and output nodes. In contrast, a RNN has feedback connections, allowing it to remember information over time. At each time-step the RNN receives a new input and produces an output based on both the current

input, and information from the previous time-steps. During training of a RNN using back-propagation-through-time, the recurrent connections are 'unrolled' to create a very deep feed-forward network [42], as shown in Fig. 1. Given the unrolled network, the lateral connections can be seen to act as memory, allowing information to flow between a potentially indefinite number of time-steps. It is commonly accepted that the performance of deep networks is due to hierarchical feature extraction that takes place over many layers [17], therefore we use a CNN to pre-process each input image into a higher-level representation before the recurrent layer.

As video re-identification involves recognising a person from a time-series of images, the use of recurrent connections may help to improve re-identification performance by allowing information to be passed between time-steps. By incorporating recurrent connections between the CNN and temporal pooling layers, we aim to better capture temporal information present in the video sequence.

Three different recurrent networks architectures are studied as candidates for our recurrent layer: Long Short Term Memory (LSTM) [61], Gated Recurrent Unit (GRU) [62] and our proposed RNN layer with residual connections (RNN-r). We also considered variants of the LSTM and GRU with residual connections [63]. The use of residual connections has shown promise in both conventional feed-forward networks [14] and with recurrent networks [63].

As described in Section III-B, $f^{(t)}$ is the vectorized output of the CNN's final layer activation maps, for the image $s^{(t)}$ observed at time t. Our proposed RNN layer is defined as follows:

$$o^{(t)} = W_i f^{(t)} + W_s r^{(t-1)} \quad (1)$$

$$r^{(t)} = Tanh(o^{(t)}) \quad (2)$$

The output, $o^{(t)} \in \mathbb{R}^{e\times 1}$, at each time-step is a linear combination of the vectors, $f^{(t)} \in \mathbb{R}^{N\times 1}$, containing information on the current input image, and, $r^{(t-1)} \in \mathbb{R}^{e\times 1}$, containing information on the RNN's state at the previous time-step. The output is computed using the fully-connected layers, $W_i \in \mathbb{R}^{e\times N}$ and $W_s \in \mathbb{R}^{e\times e}$, respectively, where e is the dimensionality of the feature embedding-space, and N is the dimension of the vectorised representation of the CNN's final layer activation maps. Note that the parameter matrix W_i is non-square, meaning that the CNN's final-layer activation maps are projected to a vector in a lower-dimensional feature embedding space. The RNN state, $r^{(t)}$, is initialised to the zero-vector during the first time-step, $r^{(0)}$, and between time-steps is passed through the Tanh non-linear function.

We note that our proposed RNN layer has some similarities to residual networks. There is a direct path, with no non-linearities, only a single linear layer, for information to flow from the CNN's final layer activation maps, through the temporal-pooling layer, to the feature representation used for re-identification. The recurrent connections are only used to modify this information flow additively, as appropriate.

In general, at each time-step the RNN produces two outputs: the vector $r^{(t)} \in \mathbb{R}^{e\times 1}$, containing the RNN's current state, which will be used during the next time-step, and, $o^{(t)}$, the RNN's output at time t, which is passed to the temporal-pooling layer. Note that Dropout [51] is used between the CNN and the recurrent layer in order to reduce over-fitting.

D. Temporal Pooling

As a final step, our re-identification architecture adds a temporal pooling layer. The purpose of this layer is two-fold. Firstly, it allows for the aggregation of information across all time steps, thus avoiding bias towards later time-steps [54], [21]. This is specially relevant in re-identification, since it may reduce the RNN's effectiveness when used to summarise the relevant information over a full sequence. Discriminative frames may appear anywhere in the sequence, not just near the end, in particular if occlusions or partial views appear when the target leaves the scene. Secondly, the temporal pooling layer aims to capture longer-term information present in the sequence to the one encoded in the optical flow or recorded by the RNN.

In the temporal pooling layer, after forward propagation of a sequence of images, the appearance features produced by the combined CNN and recurrent layer for all time-steps, $\{o^{(1)} \dots o^{(T)}\}$, are aggregated to give a single feature representing the whole sequence. We propose two approaches to temporal pooling: In the first, mean-pooling is used over the temporal dimension to produce a single feature vector v representing the person's appearance averaged over the whole input sequence, as follows:

$$v_{\mathbf{s}} = \frac{1}{T}\sum_{t=1}^{T} o^{(t)} \quad (3)$$

In the second, max-pooling over the temporal dimension is used to select the maximum activation of each element of the appearance feature vector:

$$v_{\mathbf{s}}^{i} = \max([o^{(1),i}, o^{(2),i}, ..., o^{(T),i}]) \quad (4)$$

where $v_{\mathbf{s}}^{i}$ is the i'th element of the vector $v_{\mathbf{s}}$ and $[o^{(1),i}, o^{(2),i}, ..., o^{(T),i}]$ are i'th elements of the appearance vector across the temporal dimension. We now write the complete feature extraction network as a function $R(\mathbf{s}) = v_{\mathbf{s}}$, that takes as input a time-series of person images, $\mathbf{s}$, and produces a feature vector $v_{\mathbf{s}}$ as output, representing the person's appearance over the whole input sequence. This architecture allows sequences of arbitrary length to be compared by comparing each sequence's feature vector, rather than comparing the individual images at each time-step.

E. Training Strategy

In this section we explain how the previously described network can be trained to act as a feature extractor, suitable for re-identification and wide area tracking.

1) Metric Learning: The proposed network is trained to act as a feature extractor using the Siamese network architecture [12]. The Siamese network architecture consists of two sub-networks with identical weights. When the network is presented with a pair of inputs, the sub-networks map the pair of inputs to a pair of feature vectors, which are

then compared using Euclidean distance. During training the Siamese network is shown similar and dissimilar input pairs, and it must learn to map those inputs to a feature space where similar inputs are close and dissimilar inputs are separated by a margin. Concretely, for video-based person re-identification we would like to map image-sequences from the same person to feature vectors that are close, and map sequences from different people to feature vectors that are widely separated.

Given a pair of sequences $(\mathbf{s}_i, \mathbf{s}_j)$, where each sequence has been processed using the feature extraction network to give sequence feature vectors, $v_i = R(\mathbf{s}_i)$ and $v_j = R(\mathbf{s}_j)$, we can write the Siamese network training objective as a function of the feature vectors v_i and v_j as follows:

$$E(v_i, v_j) = \begin{cases} \frac{1}{2}||v_i - v_j||^2 & i = j \\ \frac{1}{2}[\max(m - ||v_i - v_j||, 0)]^2 & i \neq j \end{cases} \quad (5)$$

where $||v_i - v_j||^2$ is the Euclidean distance between the feature vectors. When the sequences are from the same person i.e., $i = j$, the objective encourages the features v_i and v_j to be close, as measured by Euclidean distance, while for sequences from different persons i.e., $i \neq j$, the objective encourages the features to be separated by a margin m. During testing, features can be extracted for novel sequences, not observed during training, and whose identity is new and unknown, and these features can be compared using Euclidean distance, where a lower Euclidean distance indicates the sequences are more similar.

2) Joint Identification and Verification: Similar to the approach suggested in [53] for face recognition and in [41], [3] for full-body re-identification, we train the feature extraction network to satisfy both the Siamese objective and to predict the person's identity. Using the sequence feature vector, v, output by the feature extraction network, R, we can predict the identity of the person in the sequence using the standard cross-entropy loss, or softmax function, which is defined as follows:

$$I(v) = P(q = c|v) = \frac{\exp(W_c\, v)}{\sum_k \exp(W_k\, v)} \quad (6)$$

where there are a total of K identities, q is the identity of the person, and W_c and W_k refer to the c^{th} and k^{th} column of W, the softmax weight matrix, respectively. As an aside, we have found that jointly training for identification and Siamese cost is crucial for convergence. We can now define the overall training objective Q for a single pair of sequences, which jointly optimizes the Siamese cost and the identification cost as follows:

$$Q(\mathbf{s}_1, \mathbf{s}_2) = E(R(\mathbf{s}_1), R(\mathbf{s}_2)) + I(R(\mathbf{s}_1)) + I(R(\mathbf{s}_2)) \quad (7)$$

Where taking a similar approach to [53], we weight the identification cost and Siamese cost equally. The above network can be trained end-to-end using back-propagation-through-time (details of our training parameters can be found in section V). During training with back propagation through time, all recurrent connections are unrolled to create a deep feed-forward graph, where the weights of the recurrent layer and CNN are shared between all time-steps [42]. After training we discard the Siamese and identification cost functions and retain $R()$ for use as a feature extractor, where the feature vectors extracted by $R()$ can be directly compared using Euclidean distance.

IV. Wide Area Tracking Framework

A multi-target tracking algorithm based on network flow and linear programming, ELP [39], is used as baseline to evaluate the performance of our video re-identification algorithm within a conventional tracking framework. ELP is a two-stage tracking framework, where individual detections $d_k \in D$ are merged to create tracklets $\tau_i \in T$ in the first stage, and then tracklets $T = [\tau_1 ... \tau_i ... \tau_N]$ are merged in the second stage to deal with occlusions, gaps, and other detection problems. In this formulation, a tracklet is defined as an ordered set of detections $\tau_i = [d_1 ... d_k ... d_K]$, and the goal of the ELP tracker is to find the optimal set of N tracklets $T^* = [\tau_1 ... \tau_N]$ which best explain the detections i.e. the set of tracklets with the maximum posterior probability given the detection set:

$$T^* = \arg\max_T P(T|D) = \arg\max_T \prod_k P(d_k|T) \prod_i P(\tau_i) \quad (8)$$

The ELP tracker formulation translates Eq. 8 into an equivalent minimum-cost network-flow problem, where detections, in the first ELP stage, and tracklets in the second ELP stage, are modelled as graph nodes, and the cost of associating nodes i ad j is modelled as graph edges with associated costs $C_{i,j}$. Thus, the optimal set of tracklets can be found by solving:

$$T^* = \arg\max_T \sum_i C_i f_i + \sum_i C_{i,j} f_{i,j} + \sum_i C_{s,i} f_{s,i} + \sum_i C_{i,r} f_{i,r} \quad (9)$$

where s and r are the source and sink vertexes of the graph and $f_{i,j} \in [0, 1]$ are flags along the edges that allow introducing constraints to ensure valid tracking solutions, such as mutual exclusion at the starting and ending detections of each tracklet:

$$f_{s,i} + f_i \leq 1 \qquad f_{i,r} + f_i \leq 1 \quad (10)$$

and enforce conservation of flow at each detection:

$$f_{s,i} + f_i = \sum_j f_{i,j} \qquad f_{i,r} + f_i = \sum_j f_{j,i} \quad (11)$$

Two modifications are performed to the original ELP tracking system. Firstly, the cost function of the second stage is modified to use primarily the dissimilarity between the video re-identification feature vectors generated by our network v_s associated to tracklet τ_s. This modification aims to validate the performance of our features for tracklet association within a single camera field of view c. Given two tracklets τ_i^c and τ_j^c that have been generated by the first stage and observed by the same camera c, their feature vectors v_i and v_j are generated by feeding the bounding boxes associated to the tracklets into our video re-identification network. Thus, the linking cost between both tracklets is measured by the Euclidean distance between their corresponding feature vectors:

$$E(v_i, v_j) = ||v_i - v_j||^2 \quad (12)$$

Secondly, we add a third stage to the ELP tracker that allows tracklets to be linked between non-overlapping cameras. To achieve this aim, tracklets observed by different cameras are modelled as nodes of a new graph and a third minimum-cost network flow problem is solved. Given two tracklets $\tau_i^{c_1}$ and $\tau_j^{c_2}$ that have been generated by the first stage and observed by different cameras c_1 and c_2, the association cost $C_{i,j}$ can also be computed by the Euclidean distance between their respective feature vectors v_i and v_j.

However, unlike the standard re-identification problem formulation, where all probe sequences have a corresponding gallery sequence, wide area tracking is an open-world problem. This means it is possible that the probe sequence will have no correspondence in the gallery and vice versa. To cope with this possibility two extra terms are added to the cost function. Firstly, tracklets will not be associated if the time gap between them $\Delta T_{i,j}$ is larger than a temporal threshold $\Delta\tilde{T}_{max}$. Secondly, association between tracklets from different cameras will only be considered if their appearance similarity is smaller than a confidence threshold E_{max}. As a result, the final cost functions can be formulated using the linear programming notation as follows for the second stage:

$$C_{i,j}^{2nd} = C(E(v_i, v_j), 1) + C(\Delta T_{i,j}, \Delta\tilde{T}_{max}) \quad (13)$$

and for the third stage:

$$C_{i,j}^{3rd} = C(E(v_i, v_j), E_{max}) + C(\Delta T_{i,j}, \Delta\tilde{T}_{max}) \quad (14)$$

being

$$C(x, y,) = 1 - \exp - \sqrt{\frac{x}{y}} \quad (15)$$

Since no modification is performed to the ELP first stage, the cost function to link detections into tracklets is kept the same as in the original proposal [39]:

$$C_{i,j}^{1st} = C(E(d_i, d_j), D_{max}) + C(\Delta T_{i,j} - 1, \Delta T_{max}) \quad (16)$$

V. Experiments

A. Video Re-identification

In this section we evaluate our approach to video re-identification on two different datasets: iLIDS-VID [59] and PRID-2011 [18]. The iLIDS-VID dataset contains 300 persons, where each person is represented by two video sequences captured by non-overlapping cameras. The sequences range in length from 23 to 192 frames. The PRID-2011 dataset contains 749 persons, captured by two non-overlapping cameras, with sequences lengths of 5 to 675 frames. Following the protocol used in [59], we only consider the first 200 persons, who appear in both cameras.

For these experiments each dataset was randomly split into 50% of persons for training and 50% of persons for testing. All experiments were repeated 10 times with different test/train splits and the results averaged to ensure stable results. The hyper-parameters of the convolutional network were set to the same values as in [38], [41], optimised for single-shot re-identification on the Viper re-identification dataset [11]. And based on [38], the margin in the Siamese cost function was set to 2, and the feature embedding-space dimension was set to 128, which is equal to the resulting dimensionality given by the last convolutional layer. The network was trained for 500 epochs using stochastic gradient descent with a learning rate of 1e-3, and a batch size of one, alternating between showing the Siamese network positive and negative sequence pairs. A full epoch consisted of showing all positive sequence pairs and an equal number of negative pairs, random sampled from all training persons.

Given 150 persons with a maximum sequence length of 192 frames, training for 500 epochs takes approximately one day using an Nvidia GTX-980 GPU. Re-identification can then be performed efficiently, as only the new sequence must be passed through the network to produce a feature vector. Passing a single image through our network to produce a feature vector takes 0.0012s i.e. ∼830 images per second can be processed. The appearance of a person after their full video sequence has been processed by the network is represented by a single vector of length 128. This means a large gallery can be stored with a small memory footprint.

Pre-computed feature vectors are stored for all gallery sequences and can be efficient compared with the new sequence (i.e. a new vector of size 128) using a single matrix vector product. Assuming a gallery of 1000 persons, stored as a matrix of size 1000x128, comparison of a new feature vector with the gallery requires only 0.1ms.

Positive and negative sequence pairs consist of two full sequences of arbitrary length from different cameras, showing the same person or different persons respectively. During training, sub-sequences of $k = 16$ consecutive frames were used for computational reasons, where a different subset of 16 consecutive frames over the full sequence length was randomly selected at each epoch. During testing we consider the first camera as the probe and the second camera as the gallery, as in [59].

Data augmentation in the form of cropping and mirroring was applied to increase the diversity of the training sequences, and for a given sequence the same augmentation was applied to all frames during each presentation to the network. During testing data augmentation was also applied to the probe and gallery sequences, and the similarity scores between sequences averaged over all the augmentation conditions, as in [22].

1) Feature Type and Recurrent Connections: In this experiment we investigate some of the main architectural choices of our proposed system: the use of recurrent connections, and the choice of input channels. Training and testing of the network was performed with recurrent connections either disabled or enabled, and with either colour features only, or colour and optical flow features together. The results of this experiment are presented in Fig. 3 as CMC curves for the iLIDS-VID and PRID-2011 datasets.

The results show that the use of recurrent connections improves performance on both datasets regardless of the features types used, compared to the network without recurrent connections. For both datasets the best performance occurs when recurrent connections are enabled, and optical flow and colour features are used together. Performance is lowest for both datasets when recurrent connections are disabled and colour features are used alone. This suggests that our choice

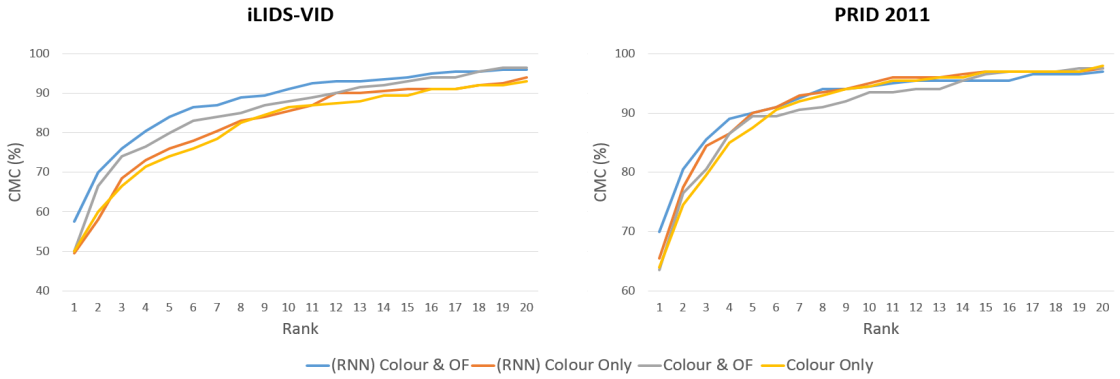

Fig. 3. CMC curves for iLIDS-VID and PRID-2011 datasets, comparing the network trained and tested on with/without recurrent connections, and with colour and optical flow input, or colour input only. Note, the vertical axis in each figure have different scales.

to explicitly embed short term and medium term temporal information into the network architecture through the use of optical flow and a recurrent layer respectively, improves re-identification performance. For the iLIDS-VID dataset this benefit is more obvious, as there is a clear separation between the performance of different methods, while for PRID-2011 dataset the performance tends to be similar, as well as very high, after rank five. Qualitative examination of the data suggests that the iLIDS-VID dataset has more cluttered backgrounds and occlusion, showing a higher complexity than PRID-2011, where the subjects are more distinct. This lower complexity may explain why all variants of our proposed method perform similarly on the PRID-2011 dataset after the candidates with similar appearance, who are more likely to be confused, are grouped together in the first five ranks and upwards.

2) Temporal Pooling: In section III-D we proposed two methods for temporal-pooling of appearance information over a sequence to give a representation of the sequence as a single feature vector: mean-pooling and max-pooling.

In this experiment we compare re-identification performance when the network has been trained and tested with either mean-pooling or max-pooling, and with the recurrent connections disabled to make the effect of the different pooling methods clearer. We also consider a baseline method [38] for computing a similarity-score between sequences that processes each frame individually using a single-frame CNN trained using a Siamese architecture and whose individual frame outputs are combined into a single decision without mean-pooling: The similarity between the sequences is then taken as the average Euclidean distance between corresponding frames. This single-shot CNN is exposed to all the data from the video sequences available in training, and trained using pairs of still-images, rather than sequence pairs, where a different single frame over the full sequence length was randomly selected at each epoch. In this experiment training and testing was carried out using the iLIDS-VID dataset.

The CMC curves of the two pooling methods and the baseline approach are shown in Fig. 4. It can be seen that mean-pooling performs better than both max-pooling and the baseline method. These results are interesting as they show that using mean-pooling to represent the whole sequence as a single feature vector leads to better performance than the baseline method which considers each frame individually. They also shows the utility of considering all the time steps equally important in the decision by using mean pooling, as opposed to max-pooling where only the feature value in the temporal step with the largest activation is employed. These results suggest that using mean-pooling over the temporal sequence of features may allow the network to better cope with noise and/or occlusions, and produces a single robust feature vector to compress and represent the person's appearance over a period of time.

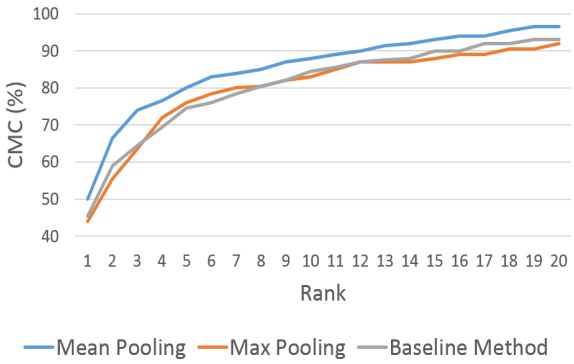

Fig. 4. CMC curves comparing different methods of computing the similarity between sequences. Two temporal pooling architectures, mean-pooling and max-pooling, are compared with a baseline method without temporal pooling.

3) Recurrent Network Type: In this section the choice of recurrent network architecture is explored. Six different recurrent networks were tested: our proposed RNN with residual connections (RNN-r), a linear layer with no recurrent connections (no RNN), Long Short Term Memory (LSTM), Long Short Term Memory with a residual connection (LSTM-r), a Gated Recurrent Unit (GRU) and a Gated Recurrent Unit with residual connection (GRU-r).

Three-fold cross-validation was used as the evaluation strategy on the two video re-identification datasets at two different learning rates, 10^{-2} and 10^{-3}, given the importance of this parameter in the recurrent network performance. The

performance is measured using both the rank 1 and the area under the CMC curve (AUC). Results are averaged first over cross validation folds and datasets.

Fig. 5 shows the comparative results between different recurrent networks. It is important to note the crucial role of the learning rate in achieving good performance. Contrary to what has been reported for other applications [47], [10], we find that the conventional LSTM did not give improved performance compared with no using RNN or our RNN-r. This suggests that the LSTM was not able to properly extract relevant temporal information for video re-identification and the increased number of parameters to be learned in this architecture damages the performance. The GRU shows inconsistent results and dependency on the learning rate.

We find that adding residual connections to the LSTM and GRU improves their performance. The LSTM and GRU with residual connections show a similar performance to the standard RNN, and give better results than not using a recurrent layer. Overall the GRU with residual connections gives the best performance, although the difference in performance compared to our standard RNN architecture is small.

We hypothesise that the use of residual connections improves performance by making it easier to train the network to perform re-identification. This is because the residual connections provide a direct path for information to flow from the CNN to the feature representation. Therefore the recurrent layer with residual connections only needs to learn how to modify the feature representation with temporal information where appropriate. This contrasts with a more conventional recurrent architecture where all information must pass through the recurrent layer. In this case the recurrent layer must learn to both faithfully represent the appearance information at its output, while also augmenting it with appropriate temporal information. Architectural modifications to make learning easier could be especially important for re-identification given that there is relatively limited data available for training.

4) Probe and Gallery Sequence Lengths: In this experiment we investigate how re-identification accuracy varies depending on the lengths of the probe and gallery sequences during the test phase, assuming a pre-trained network. Testing was performed on the iLIDS-VID dataset, and the lengths of the probe and gallery sequences were varied between 1 and 128 frames, in steps corresponding with the powers-of-two. Training lengths were fixed to 16 time steps as indicated at the start of this section. For some cases, where the desired gallery or probe length is greater than the real sequence length, we simply use the whole sequence. Probe sequences of length k are taken from the first k frames of the sequence recorded by first camera, and the gallery sequences of length k are taken from the last k frames of the sequence recorded by the second camera, since those are the farther temporal instants respectively.

Fig. 6 shows the achieved improvement in accuracy when the sum of both sequences' length increases. The thick blue line represent the improvement when both sequences are lengthened symmetrically, i.e. when the ratio between the length of probe and gallery sequences is 1. The descending thin colour lines represent the improvements obtained when the relative training to testing length ratios are varied from 2:1 to 64:1, resulting in more unbalanced probe and gallery sequence lengths. Each colour line is the result of averaging the two possible ratios X:1 and 1:X.

The results show that increasing either the probe or gallery sequence lengths improves re-identification accuracy. This accuracy increases logarithmically with the number of images. This conclusion matches the logical intuition that using a larger number of samples for each person results in an improvement in re-identification accuracy. However, this improvement is not only due to the larger amount of available images, since having similar length sequences also plays a crucial role in performance. Increasing both sequence lengths simultaneously gives the greatest improvement in accuracy, as it can be seen by the thick blue line in Fig. 6, while only increasing the length of the longer sequence has progressively less of an effect, as shown by the thin colour lines. The bigger the unbalance between the sequences compared, the smaller the improvement from using a single longer sequence, which moves from logarithmic towards linear improvement.

Results are also reported in Fig. 7 as a matrix showing the rank 1 re-identification accuracy as a function of the probe and gallery sequence lengths. It can be observed that, when different sample lengths are used, there seems to be approximate symmetry in performance when increasing either the probe sequence length or the gallery sequence length, with a slight benefit to having longer gallery sequences than probe sequences. This could prove useful for practical applications where it may be easier to collect large amounts of gallery data but where only a short probe sequence is available. When only one sample is available for each person in the gallery, increasing the probe length does not significantly improve accuracy, while if only one sample is available for the probe, increasing gallery length has a much greater effect on accuracy. This is of particular interest for those applications, such as watch-lists, where image to video re-identification is desired.

5) Comparison with the state of the art: We now compare the performance of our proposed video-based re-identification system against state-of-art methods from the literature. We also include results for the baseline DNN [38], described in Section V-A2, to put our results in context and to measure the improvement when using temporal information, as in our proposed network architecture. To ensure a fair comparison, the baseline system was trained and tested using the same datasets and same test/training split as the video-based system.

In Table I we compare the CMC results for our system, trained and tested on the iLIDS-VID and PRID-2011 datasets, with other state-of-the-art video re-identification systems. Comparing the CMC results of our proposed system with the baseline (still image based) system we can see that the video re-identification system performs better for both datasets. When we compare our results with the literature, our system shows superior performance against most other video re-identification systems, including those based on similar RNN architectures. Only AMOC [34] shows a superior performance, which is understandable since they use our system [40] as baseline. Their great improvement is achieved by using a

a) Learning rate: 10^{-2} b) Learning rate: 10^{-3}

Fig. 5. Comparison of the different RNN architectures for re-identification accuracy, where results have been averaged over i-LIDS-VID and PRID2011. Top: Area under the curve (AUC). Bottom: Rank 1 CMC.. Results are separated by learning-rate.

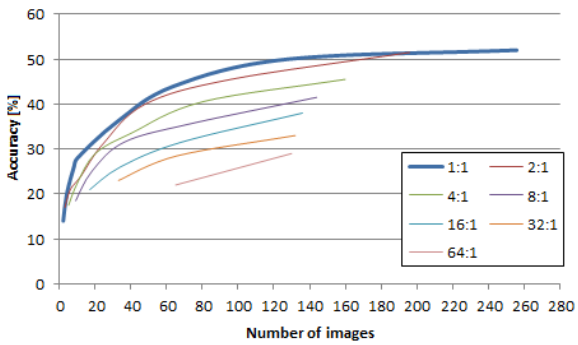

Fig. 6. Accuracy improvement in video re-identification versus the length of the compared sequences. The coloured lines represent different ratios of probe to gallery sequence length.

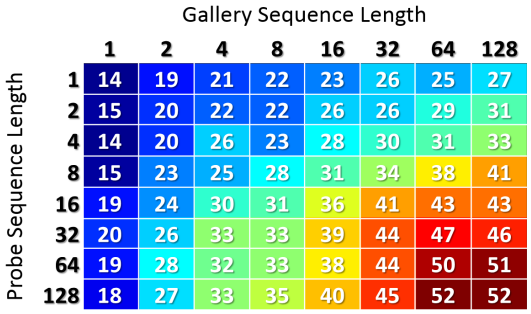

Fig. 7. iLIDS-VID rank 1 CMC re-identification accuracy as the lengths of the probe and gallery sequences are varied.

better (but offline) optical flow algorithm and a two-stream architecture where motion and appearance features are better modelled by learning them separately, fusing them before the RNN layer. Superior performance in PRID, but not in ILIDS, is also achieved in [37] and [66] by combining their unsupervised features with another state of art supervised features and using the novel XQDA metric learning method.

Furthermore, a final comparison was done using the recently proposed MARS dataset [66], which contains 1261 identities and around 20,000 tracklets. We follow the experimental setup proposed in [66], and our results are shown in Table II. For comparison purposes, we also tested a version of our system combined with XQDA, and a version of our system using a CNN pretrained on a large number of single-shot reid datasets (see last paragraph in Section V-B2 for details). It can be seen that [66] surpasses our system mainly due to a better metric learning, i.e. XQDA. Their bigger network size may also be able to better utilise the larger MARS dataset.

6) Cross-Dataset Testing: Cross-dataset testing may be a better way to estimate a system's real-world performance than evaluating performance on the same dataset used for training, which may lead to over-fitting to a particular scenario. This is due to dataset bias [38], [56], which is a form of over-fitting where the performance of a machine-learning based system,

Dataset	PRID-2011				iLIDS-VID			
CMC Rank	1	5	10	20	1	5	10	20
Ours	70	90	95	97	58	84	91	96
Baseline	55	85	94	97	38	62	71	79
LSTM [61]	47.8	77.4	90.7	94.6	41.6	70.2	86.4	92.3
LSTM+KISSME [61]	69	88.4	93.2	96.4	46.1	76.8	89.7	95.6
GRU [62]	49.8	77.4	90.7	94.6	42.6	70.2	86.4	92.3
RFA [64]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
AMOC [34]	83.7	98.3	99.4	100	68.7	94.3	98.3	99.3
CNN+Euc [66]	58.2	82.7	90.6	98.2	40.5	70.0	78.9	84.7
CNN+XQDA [66]	74.8	92.1	95.7	99.1	51.3	79.1	87.2	94.3
STA [35]	64	87	90	92	44	72	84	92
VR [59]	42	65	78	89	35	57	68	78
SRID [25]	35	59	70	80	25	45	56	66
AFDA [33]	43	73	85	92	38	63	73	82
DTDL [26]	41	70	78	86	26	48	57	69
DVR [37]	41.7	67.1	79.4	90.1	31.5	62.1	72.8	82.4
XQDA+DVR [37]	77.4	93.9	97.0	99.4	51.1	75.7	83.9	90.5

TABLE I
COMPARISON OF OUR PROPOSED APPROACH WITH THE LITERATURE ON ILIDS-VI AND PRID-2011 IN TERMS OF RANK CMC (%). METHODS 3 TO 7 ARE BASED ON RNN ARCHITECTURES.

Dataset	MARS-2106			
CMC Rank	1	5	10	20
Ours	43	61	67	73
Ours+XQDA	52	67	73	77
Ours Pretrained CNN	51	64	68	72
Ours Pretrained CNN+XQDA	56	69	73	77
CNN+Euc [66]	58.7	77.1	-	86.8
CNN+XQDA [66]	65.3	82.0	-	89.0

TABLE II
COMPARISON OF OUR PROPOSED APPROACH WITH THE LITERATURE ON MARS IN TERMS OF RANK CMC (%).

trained on a particular dataset, is much worse when evaluated on a different dataset. One cause of this problem is that any given dataset represents only a small fraction of all real-world data, making it difficult for the system to learn which aspects of the training data are essential to the problem, and which are just artefacts of the dataset.

System	Trained On	1	5	10	20
Ours	iLIDS-VID	28	57	69	81
Ours*	iLIDS-VID	14	38	51	70
Ours	MARS	18	46	61	74
Baseline	Viper	17	36	48	68
Baseline*	Viper	14	31	45	61
CD [23]*	Shinpuhkan 2014	17	-	43	52
CNN+Euc [66]	MARS	7.6	24.6	39.0	51.8

TABLE III
CROSS-DATASET TESTING ACCURACY TESTED ON PRID 2011 IN TERMS OF RANK CMC (%), WHERE * INDICATES ONLY ONE IMAGE WAS USED FOR GALLERY AND PROBE I.E. SINGLE-SHOT RE-IDENTIFICATION.

Therefore to better understand how well our proposed system generalises, we also perform cross-dataset testing, where the large and diverse iLIDS-VID and MARS datasets were used for training, and testing was performed on 50% of the PRID 2011 dataset, so that the results of this experiment can be compared with the results in Section V-A5. We also include results for the baseline system comparison trained on the Viper dataset (for details of the baseline system please see Section V-A2) as one of the most popular still-image re-identification datasets to provide a context for the results and to provide a reference point for single-shot re-identification systems in the cross-dataset setting. Testing was performed either using both the full sequences available, and to facilitate fair comparison with the literature, using a single still-image for both the probe and gallery for each person.

We can compare the results in the cross-dataset scenario with those in Table III, when the system was trained and tested on PRID 2011 dataset. The results in the cross-dataset scenario are worse, as expected, probably due to dataset bias. However it should be noted that the rank 1 performance is not much below [25] (see Table I), and is well above other single-shot re-identification systems, such as [23], even those specifically trained in PRID, such as [19] with a rank 1 CMC scores of 28. It can also be noticed there is a 100% improvement when using video re-identification that includes temporal information, which shows that our architecture is exploiting this temporal information to achieve better performance than the baseline. We include these results in the hope that others will also perform cross-dataset testing and improve the generalisation performance of re-identification systems. The best performing method in the previous section AMOC [34] did not report result in a cross-dataset set-up in order to evaluate if previous performance is due to over-fitting or real improvement. The system presented in [66], which reported better results in MARS and PRID-2011 in the previous section, exhibits worse results than our system in the cross-dataset setting, where no training or fine tuning of the CNN, nor additional metric learning is allowed to be used in the testing dataset.

B. Video Re-identification for Wide Area Tracking

In this section we validate our re-identification approach with a realistic experiment in wide area tracking. To do this we firstly integrate our video re-identification system with an existing multi-target tracking system in order to validate the use of video re-identification features for linking tracklets within the same camera. Secondly, we evaluate the use of

video re-identification features for person tracking over a network of non-overlapping cameras with unknown layout.

To ensure that our results are robust and reflect the system's expected real-world behaviour, the dataset used for testing wide-area tracking performance is completely different from the datasets used for training the re-identification system. This means that these experiments are essentially performed in the cross-dataset setting, which is similar to how a wide-area tracking system deployed in a real scenario would be used.

In these experiments on wide area tracking, our video re-identification system was trained using both the iLIDS-VID and PRID 2011 datasets (500 people), with colour and optical flow inputs. Note that unlike previous experiments, all persons from both datasets were used for training the system, giving around 480 training persons.

As baseline, we compare the performance of our re-identification features within the wide-area tracking framework against the same system using colour histograms h features. The cost function for calculating the dissimilarity between the colour histogram features for tracklets i and j using Bhattacharyya distance is defined as:

$$E(h_i, h_j) = 1 - \sum \sqrt{h_i \cdot h_j} \tag{17}$$

1) Within-camera re-identification: To test within camera tracking performance the ELP-tracker was used [39]. This tracker was modified so that either colour histogram or re-identification features could be used for calculating the appearance similarity between tracklets. All the other tracking system's parameters were kept constant. The tracker's performance with either feature type was evaluated in the MOTChallenge (2D MOT 2015) training and validation set [30], and the MOTA was used as main evaluation metric.

Table IV shows the tracking results for both the colour histogram based ELP tracker and ELP tracker using re-identification features, in each dataset as well as the global average. The parameter *gap*, is defined as the maximum temporal threshold $\Delta\tilde{T}_{max}$ allowed when linking tracklets. This parameter was varied between 1 and 5 seconds, which gives the tracker a greater opportunity to correctly link tracklets at the cost of also increasing the chances of making a mistake.

We note that the video re-identification features are as effective as the conventional colour histogram features, for successfully link tracklets. Although there is no significant performance difference when tracklet linking is constrained to the most obvious links (*gap*=1), video re-identification features shows better performance and less tendency to introduce errors when the linking constraints are relaxed, as well as better consistency independently of the chosen parameters and the gap threshold.

2) Inter-camera re-identification: Two experiments were performed to compare the performance of video re-identification features for linking tracklets between non-overlapping cameras in a realistic scenario. This differs from standard person re-identification experiments as there are out-of-sample persons i.e., this is an open-world scenario, making the association task more difficult.

A two minute sequence from the DukeMTMC [46] dataset training and validation-set [30] was first chosen. The sequence contained a total of 51 people, where 10 people transited between cameras. The ELP-tracker was used to track all persons in camera 5 and camera 2, producing a set of tracklets for each camera. The goal of this experiment was to correctly associate the tracklets of persons appearing in both cameras. Note that the re-identification system did not have prior knowledge of which persons appeared in both cameras, and which persons only appeared in one camera.

As in the previous experiment, our video re-identification system, trained on the complete iLIDS-VID and PRID 2011 datasets, was compared against a baseline re-identification using colour histogram features. A threshold E_{max} was used to prevent mismatches, and the threshold value was set using a preliminary experiment on the MOTChallenge dataset. Given the different similarity scales used by video re-identification features and normalised colour histograms, the optimal resulting threshold was set to $E_{max} = 5$ for the former and to $E_{max} = 0.1$ for the later.

In order to test the importance of the sequence length for the re-identification between cameras, we additionally varied the number of frames that were used to calculate the appearance features. For video re-identification features this parameter controls how many video frames were passed to the recurrent network. For colour histogram features we calculated the average colour histogram features over the specified number of time-steps.

Fig. 8 shows the results of both systems as percentages normalising by the number of targets in the ground-truth and by the number of links attempted by each system. Note that in the left image the percentage can be greater than 100% as the system can attempt a number of links up to the real number of people appearing in either camera. The results show that the video re-identification features do a much better job of correctly linking tracklets between cameras, while avoiding mismatches. The colour histogram features are not capable of reliably linking tracklets between cameras. We can additionally see that using longer tracklets significantly improves the results, and it is more effective than conventional single-frame re-identification. This is to be expected as the video re-identification system can make use of additional motion information to produce better matches, as well as integrating evidence over the whole sequence.

In a second experiment, the full 50-minute long DukeMTMC training and validation sequence [46], [30], containing a total of 1233 people during the sequence (436 in camera 5 and 797 in camera 2), is used to validate the previous conclusions. Among the 1233 subjects, 271 people transit between cameras 2 and 5. The number of frames used to calculate the appearance features was fixed to a maximum of 512 frames, when available, as per the previous experiment.

Three systems were compared using the ELP tracker and different cost functions for linking tracklets between cameras in the third stage of the tracker: cost based on colour histograms, ELP + colour hist. (see Eq. 17), cost based on our video re-identification features, ELP + video reID (see Eq. 12), and finally cost based on single frame re-identification

Tracking System	Gap	ADL-Rundle6	ADL-Rundle8	ETH-Bahnhof	ETH-Pedcross2	ETH-Sunnyday	KITTI-13	KITTI-17	PETS09-S2L1
ELP + video reID	1	59.3	38.4	77.4	35	33.3	7.8	30.6	20.9
	2	60.6	38.4	77.5	34.5	33.5	7.8	31	20.6
	5	60.6	38.4	77.9	33.9	33.5	8	30.3	20.6
ELP + colour hist.	1	59.4	38.4	77.6	34.3	34	8.3	31.7	21.2
	2	59.4	39.6	77.6	33.3	35.4	8.3	31.7	20.8
	5	59.4	39.6	77.9	32.1	35.8	8.3	31.7	20.8

Tracking System	Gap	TUD-Campus	TUD-Stadtmitte	Venice-2	Average
ELP + video reID	1	27	57.8	16.2	30.1
	2	27	57.8	16.3	30.1
	5	27	61.5	15.4	29.9
ELP + colour hist.	1	27	58.4	15.4	30.2
	2	27	61.6	13.6	29.8
	5	27	61.1	12.6	29.5

TABLE IV
COMPARISON OF SINGLE CAMERA TRACKING PERFORMANCE IN THE MOTCHALLENGE DATASET USING THE ELP TRACKING FRAMEWORK. TRACKING PERFORMANCE WITH COLOUR HISTOGRAM FEATURES AND OUR VIDEO RE-IDENTIFICATION FEATURES IS COMPARED.

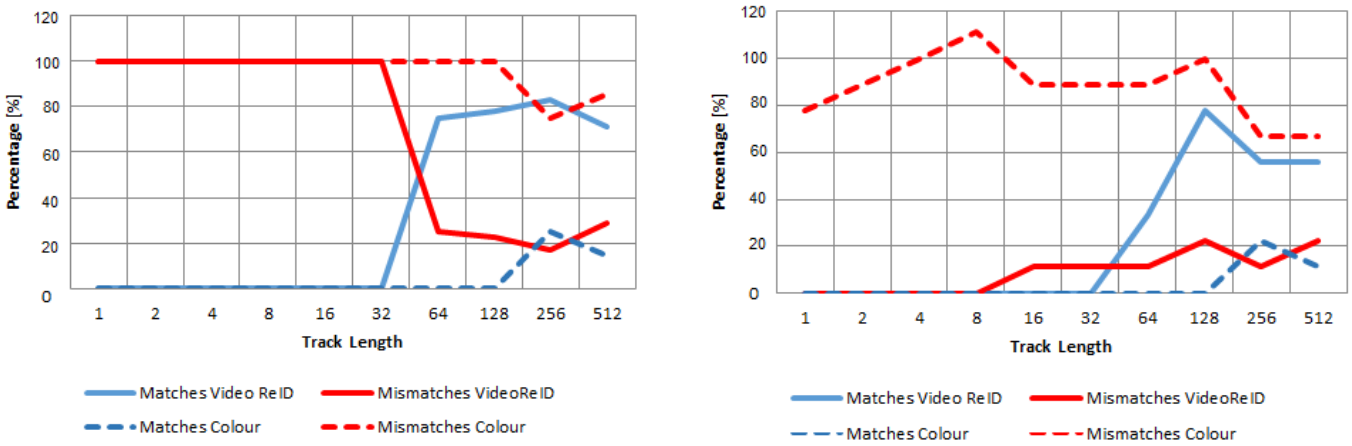

Fig. 8. Re-identification matches and mismatches for the selected two minute sequence from the DukeMTMC training and validation set. Results are shown for the system using our video re-identification features and colour histograms, with respect to (left) the total amount of targets transiting cameras, and (right) the total number of attempted links.

system, ELP + image reID. The third system is implemented by removing the RNN layer and optical flow input from our system but keeping the same CNN architecture and temporal pooling layers.

The first three rows of Table V show the results of our proposed system, using either video re-identification features or using colour histograms features in the third stage cost function of the EPL tracker. It can be seen that there is a drastic improvement in rank-1 re-identification accuracy, from 13.3% using colour histogram features, to 47.2% using image-only re-identification, and to 52.8% using our full video re-identification system. These results are especially remarkable given the length of the DukeMTMC training and validation sequence and the open-world setting.

Finally, as an indication of the potential of our wide area tracking approach, a different single image re-identification network was also tested, ELP + image reID (large training). Note that this network is not directly comparable with our proposed video re-identification approach due to the use of different training datasets. This network used the same CNN architecture but no recurrent layer. Mean pooling was used during testing to summarise the appearance features from each tracklet sequence, but was not used during training of this network. The network was trained with a total of ∼70,000 training images from ∼6000 different persons, resulting from the combination the following re-identification datasets: Viper, i-LIDS, CAVIAR4REID, 3DPeS, PRID, TownCentre, GRID, SARC3D, CUHK, CUHK03, Market1501, Raid and ETHZ. Our results with this network show that moving to a much larger training set improves the network's generalisation ability, resulting in a 19.2% performance improvement with respect to the similar single image reID network. Since this performance improvement was achieved without the advantages of adding a recurrent layer or other temporal information, we hypothesize that the future availability of significantly larger and more diverse video re-identification datasets will allow our system to further improve.

VI. CONCLUSION

In this paper we have introduced a novel temporal deep neural network architecture for video re-identification applied to wide area tracking. The use of optical flow, recurrent layers and mean-pooling allows us to embed the temporal hierarchy inherent to the re-identification problem in the form of

System	Matches	Mismatches	Linking Attempts	Total People
ELP + video reID	143	51	194	271
ELP + image reID	128	36	164	271
ELP + colour hist.	36	160	196	271
ELP + image reID (large training)	180	35	215	271

TABLE V
RE-IDENTIFICATION MATCHES AND MISMATCHES IN THE COMPLETE DUKEMTMC TRAINING AND VALIDATION SEQUENCE (CAMERAS 2 AND 5) USING: VIDEO REID FEATURES, SINGLE-IMAGE REID FEATURES AND COLOUR HISTOGRAM FEATURES. ALSO SHOWN ARE RESULTS OF A STILL-IMAGE REID SYSTEM TRAINED ON SEVERAL DATASETS.

short, middle and long term temporal information respectively. Results were first evaluated in two standard datasets under a close-world set-up, and surpass almost all other methods in the video re-identification literature except AMOC [34] in iLIDS, and AMOC [34] and XQDA-based methods [37], [66] in PRID, and all methods reporting results in a cross-dataset setting. The re-identification features extracted by the network are also integrated into a multi-target tracking framework for tracking within a non-overlapping camera network. Thus demonstrating the system's great potential for true wide area tracking and achieving an unprecedented performance in a open-world re-identification experiment. As future work, we aim to exploit larger datasets such as MARS [66], recently introduced by the scientific community, to create larger architectures able to better address the video-reidentification problem and wide-area tracking.

REFERENCES

[1] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*, pages 179–184. IEEE, 2011. 2

[2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(6):937–965, 2005. 2

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. *arXiv preprint arXiv:1607.05369*, 2016. 2, 5

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 1, page 6, 2011. 2

[5] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *ICIAP*, pages 179–189. 2009. 2

[6] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015. 2

[7] J. Fang, Q. Wang, and Y. Yuan. Part-based online tracking with geometry constraint and attention selection. *IEEE Trans. on Circuits and Systems for Video Technology*, 24(5):854–864, 2014. 1

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 2

[9] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016. 2

[10] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Proc. ICASSP*, 2013. 8

[11] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, volume 3, 2007. 1, 6

[12] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742, 2006. 2, 3, 4

[13] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ICDSC 2008*, pages 1–6, 2008. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[15] C. H. Heartwell and A. J. Lipton. Critical asset protection, perimeter monitoring and threat detection using automated video surveillance - a technology overview with case studies. In *Proceedings. International Carnahan Conference on Security Technology*, pages 87–, 2002. 1

[16] D. Held, S. Thrun, and S. Savarese. Deep learning for single-view instance recognition. *arXiv preprint arXiv:1507.08286*, 2015. 2

[17] M. Hermans and B. Schrauwen. Training and analysing deep recurrent neural networks. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 190–198. 2013. 4

[18] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. 2011. 2, 6

[19] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 10

[20] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. 2012. 2

[21] S. Hochreiter and J. Schmidhuber. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479, 1997. 4

[22] A. G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013. 6

[23] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li. Cross dataset person re-identification. In *ACCV Workshops*, pages 650–664, 2014. 10

[24] S. Karaman and A. D. Bagdanov. Identity inference: generalizing person re-identification scenarios. In *ECCV Workshops*, pages 443–452, 2012. 2

[25] S. Karanam, Y. Li, and R. Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPR Workshops*, pages 33–40, 2015. 10

[26] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015. 10

[27] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 275–1, 2008. 2

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012. 2

[29] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013. 2

[30] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr 2015. arXiv: 1504.01942. 1, 11

[31] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601. IEEE, 2013. 2

[32] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2

[33] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, 2015. 10

[34] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, and S. Yan. Video-based person re-identification with accumulative motion context. *arXiv preprint arXiv:1701.00193*, 2017. 2, 8, 10, 13

[35] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *CVPR*, pages 3810–3818, 2015. 10

[36] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981. 3

[37] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197210, 2017. 2, 9, 10, 13

[38] N. McLaughlin, J. Martinez-del Rincon, and P. Miller. Data-augmentation for reducing dataset bias in person re-identification. pages 1–6, Aug 2015. 2, 6, 7, 8, 9

[39] N. McLaughlin, J. Martinez Del Rincon, and P. Miller. Enhancing linear programming with motion modeling for multi-target tracking. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–77, 2015. 5, 6, 11

[40] N. McLaughlin, J. Martinez Del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE*

Conf. Computer Vision and Pattern Recognition (CVPR), 2016. 2, 8
[41] N. McLaughlin, J. Martinez-del Rincon, and P. Miller. Person reidentification using deep convnets with multitask learning. *IEEE Trans. on Circuits and Systems for Video Technology*, 27(3):525–539, 2017. 5, 6
[42] M. C. Mozer. A focused back-propagation algorithm for temporal pattern recognition. *Complex systems*, 3(4):349–381, 1989. 4, 5
[43] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern recognition*, 36(9):1997–2006, 2003. 2
[44] J. Nebel, M. Lewandowski, J. Thevenon, F. Martinez, and S. Velastin. Are current monocular computer vision systems for human action recognition suitable for visual surveillance applications? In *Proc. International Symposium on Visual Computing*, pages 290–299, 2011. 1
[45] M. S. Nixon. A step beyond: advances in gait and soft biometrics. *Biometric Technology Today*, 2016(10):9 – 11, 2016. 1
[46] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 11
[47] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014. 8
[48] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015. 2
[49] M. Shah, O. Javed, and K. Shafique. Automated visual surveillance in realistic scenarios. *IEEE MultiMedia*, 14, 2007. 1
[50] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2
[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3, 4
[52] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *arXiv preprint arXiv:1605.03259*, 2016. 2
[53] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 5
[54] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 4
[55] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 2
[56] A. Torralba, A. Efros, et al. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011. 9
[57] R. R. Varior, G. Wang, and J. Lu. Learning invariant color features for person re-identification. *arXiv preprint arXiv:1410.1035*, 2014. 2
[58] Q. Wang, J. Fang, and Y. Yuan. Multi-cue based tracking. *Neurocomputing*, 131:227236, 2014. 1
[59] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. 2014. 1, 2, 6, 10
[60] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009. 2
[61] L. Wu, C. Shen, and C. van den Hengel. Convolutional lstm networks for video-based personre-identification. *arXiv preprint*, 2016. 2, 4, 10
[62] L. Wu, C. Shen, and C. van den Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint*, 2016. 2, 4, 10
[63] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 4
[64] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person reidentification via recurrent feature aggregation. In *European Conference on Computer Vision (ECCV)*, page 701716, 2016. 2, 10
[65] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014. 2
[66] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884, 2016. 2, 9, 10, 13